2K

УДК 004.822, 004.5

І.А. Козак
ДВНЗ Київський національний економічний університет ім. В. Гетьмана,
м. Київ, Україна
kozachka@ukr.net

Централізований підхід до опису Web-онтологій

Досліджено існуючі підходи до створення та відображення онтологій. Запропоновано автоматичне створення єдиної бази онтологій на основі використання методів лінгвістичного розпізнавання образів.

Вступ

Останнім часом онтології набувають все більшого значення для найрізноманітніших типів інформаційних систем (ІС). Так, вони можуть використовуватися:
– в системах електронної комерції і віртуальних організацій для інформаційного обміну між учасниками (у тому числі при використанні програмних агентів);
– в інтелектуальних пошукових системах для тематичної рубрикації документів;
– в системах автоматизованого отримання знань від експертів;
– в системах автоматизації проектування та ін.

Як приклад можна навести роботи [1-3]. Про зростаючу роль онтологій для інформаційних систем свідчить також введення в роботі [4] поняття керованих онтологіями інформаційних систем (Ontology-Driven Information Systems – ODIS). З точки зору «часового виміру» онтології можуть використовуватися в ODIS під час їх розробки, використання, інтеграції і т.п. А з точки зору «структурного виміру» онтології можуть підтримувати різні частини інформаційної системи – інтерфейси користувача, бази даних чи інші компоненти.

Однак, не дивлячись на зростаючий «попит» на онтології, не існує ні єдиного загальноприйнятого визначення, що таке «онтології», ні єдиної думки про те, що вони мають включати в себе, ні єдиної методології побудови онтологій. Розробка онтологій на сьогодні носить скоріше дослідницький характер, що, звичайно ж, стримує їх використання.

Аналіз останніх досліджень

Детальні переліки та аналіз існуючих визначень поняття «онтологія» здійснено в роботах [5-6]. Ми не ставимо за мету дати ще одне визначення онтології, а лише вкажемо на ключові особливості, з якими погоджується більшість авторів: онтологія описує поняття предметної області та відношення між цими поняттями, відношення можуть бути різних типів.

Щодо методів побудови онтологій, то на сьогодні існує ряд мов, призначених для формального опису онтологій. Серед найбільш відомих і використовуваних: KIF (Knowledge Interchange Format) [7], DAML+OIL (DARPA Agent Markup Language) [8], OWL (Ontology Web Language) [9]. Існують також інструментальні засоби, що підтримують розробку онтологій відповідно до цих специфікацій.

Аналізуючи конструкції даних мов формального опису, можна помітити, що навіть в найбільш розширеній з них OWL [9], що включає OIL+DAML, існують детальні можливості лише для задання класів, підкласів та їх членів (таксономії), для інших же типів відношень не передбачаються спеціальні елементи – їх можна задавати лише через властивості класів. На практиці більшість із вже створених онтологій є максимум ієрархічною структурою понять предметної області. Тобто розробниками розглядаються лише такі відношення між поняттями, як «вид-клас» та рідше «об'єкт-атрибут».

Тоді, як в ряді робіт, серед яких можна назвати [10], розроблено класифікацію властивостей онтологій та уточнено перелік структурних властивостей (таксономічні зв'язки, композиційні зв'язки, топологічні зв'язки, зв'язки сутностей з процесами, причинно-наслідкові зв'язки, часові та просторові зв'язки).

Відповідно до концепції Semantic Web [11] створення RDF-описів та OWL-онтологій покладене на окремих розробників. І на сьогодні, попри досить незначні напрацювання в плані розробки онтологій, вже виникає проблема узгодженості онтологій, яка полягає в тому, що різними розробниками для однієї й тієї ж предметної області можуть бути створені онтології, синтаксично або семантично гетерогенні, і для їх сумісного використання необхідна трансляція або відображення (виявлення відповідності між поняттями двох онтологій) [12].

Існує декілька підходів до вирішення проблеми відображення онтологій, перший з яких – ручне відображення, шляхом встановлення відношень між концептами, здійснювалося для деяких великих онтологій. Як приклад такого підходу можна навести роботу [13]. Проблема застосування ручного відображення в тому, що розмір онтологій може бути дуже великим і продовжуватиме нарощуватися, що вимагатиме надзвичайно багато людських зусиль для їх відображення. Тому, природньо, що дослідники шукають шляхи відображати онтології автоматично.

Досить значна кількість досліджень, серед яких і [14], присвячені розробці засобів відображення онтологій на основі методів машинного навчання, серед яких особливою популярністю користуються методи класифікації текстів. Однак результати тут залежать від якості навчальних даних, а підготовка їх вручну для сотень понять досить трудоємка і дорога (хоч і не настільки, як ручне відображення), що зменшує привабливість текстової класифікації.

В деяких останніх дослідженнях, серед яких можна виділити [15], пропонується використовувати дані Веб (результати пошукових серверів) для виявлення текстових екземплярів та оцінки умовної ймовірності. Однак результати цих робіт поки що незадовільні.

Ще одним підходом до відображення онтологій є, запропоноване в [16], зіставлення імен понять на основі їх лексичної подібності та використання спеціально розроблених словників (WordNet), в яких описані відношення між концептами (синонімія) та властивості ряду концептів.

Аналіз ряду засобів відображення онтологій зроблено в роботі [17].

Альтернативним напрямком досліджень є автоматичне створення онтологій, яке на сьогодні зводиться до автоматичного анотування текстів у Web. Аналіз робіт в цьому напрямку подано в [18], де показано їх обмеження виділенням певного типу відношень для анотування, або ж використанням для анотування певної онтології. В свою чергу, автори роботи [18] пропонують використовувати для аналізу веб-сторінок середовище Ontological Semantics (OntoSem) – http://www.ontologicalsemantics.com/, що створювалось для задач автоматичного перекладу та семантичного аналізу текстів. Результати аналізу пропонується автоматично відображати у OWL – описи за допомогою OntoSem2OWL.

Таким чином, можна виділити два основних напрямки досліджень, пов'язані з онтологіями:
– дослідження онтологічних властивостей для наступного формального представлення онтологій;
– дослідження можливостей відображення онтологій.

Другий напрямок, на нашу думку, є вторинним і випливає з «розподіленого» підходу до створення онтологій, який виправданий складністю їх створення.

Мета статті. Нами ставилась задача розглянути можливості «централізованого» автоматичного створення онтологій на основі виявлення в текстах, за допомогою методів лінгвістичного розпізнавання образів, основних онтологічних властивостей предметної області.

Результати досліджень

Для реалізації експериментів було розроблено морфологічний словник та спеціалізоване програмне забезпечення, за допомогою якого здійснюється морфологічний, лексичний та синтаксичний аналіз текстів з Web (на російській мові).

Для лексичного аналізу використовуються продукційні правила знаходження дієслівних, іменних, прислівникових груп. На виході отримуємо потік лексем з морфологічними ознаками (частина мови, рід, число, відмінок).

Для виявлення онтологічних властивостей була розроблена база правил, на основі яких можливе виділення з текстів різних типів онтологічних відношень.

Наприклад, як ознаки класифікації можна виділити наявність в межах речення:
– підмета-іменника (об'єкт класифікації);
– «ключових слів» («подразделить», «различать», «классифицировать», «классификация», «классы» і т.п);
– іменної групи одного із зразків («по» + іменник (в давальному відмінку) + іменник (типу процеси, в родовому відмінку); «по» + прикметник (в давальному відмінку) + іменник (в давальному відмінку) і т.п.) – є необов'язковою класифікаційною ознакою;
– переліку іменників і/або іменних груп, через кому або кому з крапкою в одному роді і числі, які є класифікаційними групами і наявні в межах речення або в наступних абзацах та ін.

В результаті аналізу текстів виявляються онтологічні моделі, на основі яких формується єдина база онтологій різноманітних предметних областей та індекси для проаналізованих веб-сторінок.

Такий централізований підхід до створення онтологій має наступні переваги порівняно з підходом [11]:
– обробка даних в базі даних здійснюється швидше, ніж окремих текстових файлів (OWL);
– усувається необхідність узгодження розрізнених онтологій;
– забезпечується можливість аналітичної обробки всіх проіндексованих сторінок та ін.

Серед проблем створення подібної єдиної бази онтологій:
– необхідність використання певного формалізму для опису онтологій;
– повнота задання правил виявлення онтологічних моделей;
– використання значних машинних ресурсів та ін.

Як формальна модель опису онтологій нами пропонується використання «багатовимірної» семантичної моделі:

$$S = \{V_1, V_2, ..., V_n\},$$

де V_n – вимір моделі.

Кожний вимір є множиною відношень певного типу для різних об'єктів, має свою структуру і характеристики.

Вимір таксономії для і-го об'єкта за k-ю класифікаційною ознакою може бути описаний за формулою:

$$T_i^k = \left\langle O_i, A_i^k, \left\{ Z_{ij}^k \right\} \right\rangle, \qquad (1)$$

де O_i – об'єкт класифікації, A_i^k – класифікаційна ознака, $\left\{ Z_{ij}^k \right\}$ – множина j-х значень для і-го об'єкта за k-ю ознакою. Вимір композиції для і-го об'єкта:

$$P_i = \left\langle O_i, \{A_i\} \right\rangle,$$

де $\{A_i\}$ – множина атрибутів і-го об'єкта.

Вимір топології можна описати аналогічно до виміру композиції.

Зв'язки об'єктів з процесами, або процесну модель можна задати:

$$F_{ij} = \left\langle O_i, \left\{ A_{ij} \right\}, N_j, R_{ij}, U_j \right\rangle,$$

де N_j – назва функції, R_{ij} – тип функціонального зв'язку (по входу чи по виходу, м.б. інший – контроль, механізм), $\{A_{ij}\}$ – множина атрибутів об'єкта, задіяна в функції, U_j – виконавець процесу.

Для кожної функції можуть бути також задані процедури перетворення даних:

$$F_j = \left\langle N_j, \{A_j^{'}\}, \left\{ A_j^{v} \right\}, \left\{ \Phi_j \right\} \right\rangle,$$

де $\{A_j^{'}\}$ – множина вхідних атрибутів для функції j, $\{A_j^{v}\}$ – множина вихідних атрибутів функції j, $\left\{ \Phi_j \right\}$ – множина функцій перетворення вхідних атрибутів у вихідні.

Причинно-наслідкові і часові зв'язки між процесами представимо як сценарії:

$$C_j = \left\langle N_j, \left\{ N_j^a \right\}, \left\{ N_j^p \right\}, Y_j, N_{j-1}, M_l \right\rangle,$$

де N_j – функція сценарію, $\{ N_j^a \}$ – альтернативні функції, $\{ N_j^p \}$ – паралельні функції, Y_j – умова виконання, N_{j-1} – попередня функція.

Дана модель названа нами багатовимірною семантичною, оскільки кожна онтологічна модель може бути розглянута як окремий вимір єдиної моделі та задана як бінарні відношення певного типу.

В процесі розв'язання задачі автоматичної побудови онтологій ми також стикнулися з проблемою узгодження онтологій ще на етапі їх створення, оскільки одні і ті ж онтологічні відношення можуть бути описані у різних текстах по-різному.

Наприклад, для O_i об'єкта класифікації виділено деяку таксономію T_i^k. При знаходженні в тексті нової таксономії для O_i-го об'єкта потрібно встановити відповідність двох таксономій. Для цього має бути здійснена перевірка на предмет збігу елементів таксономій – класифікаційних ознак та окремих таксонів (для інших онтологічних моделей це будуть інші елементи).

У нашому випадку для такої перевірки використовуються лексичний аналіз (із використанням морфологічного словника) та вже описані в базі даних відношення (у т.ч. тотожності, що включає і синонімію) для зняття семантичної гетерогенності.

Якщо перетин множин елементів певного виду для однотипних онтологічних моделей складає 50 % і більше від загального числа елементів хоча б однієї з множин – такі моделі вважаються подібними. В протилежному випадку моделі вважаються різними.

Під цінністю елемента онтології розуміється ймовірність його використання у подібних моделях P_e, яка визначається як відношення кількості використань елемента у подібних моделях до загального числа подібних моделей.

Тобто ми інтерпретуємо цінність інформації як її використовуваність або ж унікальність. Проте для унікальної інформації на основі такого підходу не враховується її істинність.

Цінність окремої моделі визначається як нормована сума цінностей її елементів:

$$C^m = \frac{\sum\limits_{e,e=(1,E^m)} P_e^m}{E^m},$$

де E^m – кількість елементів m-ї моделі.

Актуальною вважається більш цінна онтологія. До складу актуальної онтології додатково включаються елементи, що мають цінність більшу 0,5.

Також можливе врахування «коефіцієнта довіри» до тексту, з якого виявляється модель, що може бути визначений в результаті віднесення тексту до певної категорії (наукова стаття, студентський реферат, популярна стаття і т.п.).

Принципово, наш метод узгодження онтологій подібний до описаного в [16], однак основною його відмінністю є поступовість «уточнення» онтологічних моделей, тоді як в [16] це статичний обмежений словник взаємозв’язків.

Висновки

В даній роботі пропонується автоматичне створення онтологій за допомогою спеціалізованого програмного забезпечення, здатного виявляти в тексті структурні властивості онтологій на основі лінгвістичного розпізнавання образів. Для формалізованого опису онтологій запропоновано використання багатовимірної семантичної моделі.

Подальші дослідження в даному напрямку пов’язані із:

– уточненням правил виділення онтологічних моделей з текстів предметних областей;

– удосконаленням багатовимірної семантичної моделі на основі виявлення онтологічних «образів» в текстах,

– розробкою інтерфейсу для надання можливості використання онтологій широкому колу користувачів та ін.

Звичайно, на основі запропонованого нами підходу можливе також формування OWL-описів за допомогою відповідної сервісної програми. Однак, на наш погляд, більш доцільним є створення індексів для екземплярів відношень у процесі виявлення онтологічних моделей з текстів. У цьому випадку можливе також практичне застосування запропонованого підходу для розробки пошуково-аналітичних сервісів.

Література

1. Fonseca F.T. Ontology-driven geographic information systems // Thesis. The Graduate School. – The University of Maine. – May, 2001.
2. Келеберда И.Н., Лесная Н.С., Репка В.Б. Использование мультиагентного онтологического подхода к созданию распределенных систем дистанционного обучения // Educational Technology & Society. – 2004. – № 7 (2).

3. Gribova V., Kleshchev A. From an ontology-oriented approach conception to user interface development International // Journal Information theories & applications. – 2003. – Vol. 10, № 1. – P. 87-94.
4. Guarino, N. Formal Ontology and Information Systems, in N. Guarino (Ed.) Formal Ontology in Information Systems. – Amsterdam, Netherlands: IOS Press (1998). – P. 3-15.
5. Клещев А.С., Артемьева И.Л. Математические модели онтологий предметных областей. Часть 1. Существующие подходы к определению понятия «онтология» / Научно-техническая информация. – Серия 2 «Информационные системы и процессы». – 2001. – № 2. – С. 20-27.
6. Пальчунов Д.Е. Моделирование мышления и формализация рефлексии I. Теоретико-модельная формализация онтологии и рефлексии // Философия науки. – 2006. – № 4 (31). – С. 86-114.
7. Knowledge Interchange Format, draft proposed American National Standard (dpANS) NCITS. T2/98-004. Режим доступу: http://logic.stanford.edu/kif/dpans.html
8. DAML+OIL (March 2001) reference description. – Режим доступу: http://www.w3.org/TR/daml+oil-reference
9. OWL Web Ontology Language Semantics and Abstract Syntax. W3C Recommendation 10 February 2004. – Режим доступу: http://www.w3.org/TR/2004/REC-owl-semantics-20040210/
10. Шалфеева Е.А. Классификация структурных свойств онтологий // Искусственный интеллект. – 2005. – №. 3. – С. 67-77.
11. Berners-Lee, Tim; James Hendler and Ora Lassila (May 17, 2001). The Semantic Web // Scientific American Magazine. – Режим доступу: http://www.sciam.com/article.cfm?id=the-semantic-web&print=true.
12. Dou D., Mcdermott D. Qi. P. Ontology Translation on the Semantic Web // In Proceedings of International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2003). LNCS 2888. – Berlin Heidelberg. Springer-Verlag. – 2003. – P. 952-969.
13. Niles I., Pease A. Mapping WordNet to the SUMO Ontology // Proc of the IEEE International Knowledge Engineering conference (2003).
14. Zhongli Ding, Yun Peng, Rong Pan, Yang Yu. A Bayesian Methodology towards Automatic Ontology // Proceedings of the AAAI-05 C&O Workshop on Contexts and Ontologies: Theory, Practice and Applications. – July 09, 2005.
15. Yang Yu, PengY. Learning the Semantic Meaning of a Concept from the Web. Z. Kobti and D. Wu (Eds.): Canadian AI 2007, LNAI 4509. – 2007. – P. 98-109.
16. Li J.: LOM a lexicon based ontology mapping tool. – Режим доступу: http://reliant.teknowledge.com/DAML/I3con.pdf.
17. Namyoun Choi, Il-Yeol Song,Hyoil Han. A Survey on Ontology Mapping. SIGMOD Record. – 2006. – Vol. 35, № 3. – Sep.
18. Akshay Java, Sergei Nirneburg, Marjorie McShane, Timothy Finin, Jesse English, Anupam Joshi. Using a Natural Language Understanding System to Generate Semantic Web Content // Final version to appear, International Journal on Semantic Web and Information Systems. – 2007. – 3 (4). – Режим доступу: http://www.igi-pub.com/.

И.А. Козак

Централизованный подход к описанию Web-онтологий

Исследованы существующие подходы к созданию и отображению онтологий. Предложено автоматическое создание единой базы онтологий на основе использования методов лингвистического распознавания образов.

I.A. Kozak

Site Approach to Description Web-Ontology

The existing approaches for ontology creation and image are explored. Automatic making the united ontology base with use the methods of linguistical artificial perception is offered.

Стаття надійшла до редакції 21.07.2008.